Setting Standards on Alternate Assessments

NCEO Synthesis Report 42

Published by the National Center on Educational Outcomes

Prepared by:

Ed Roeber
Measured Progress

April 2002

Executive Summary

As more and more states and districts develop their alternate assessments for those students unable to participate in the regular assessment, they are faced with the challenge of setting standards for their alternate assessments. Despite the variability in alternate assessments currently developed, including checklists, structured and unstructured observations, performance assessments, samples of student work, and portfolios, all need to be scored and assigned proficiency levels. With background information on the nature of student scores from alternate assessments and the ways in which alternate assessment results are reported, this report identifies common standards-setting techniques and how they might be applied to alternate assessments. The techniques addressed are: reasoned judgment, contrasting groups, modified Angoff, bookmarking or item mapping, body of work, and judgmental policy capturing. It is recommended that the technique selected take into account not only the technical aspects of the alternate assessment strategies, but also the practical aspects of implementing the standard-setting technique for the alternate assessment process.

Setting Standards on Alternate Assessments

The Individuals with Disabilities Education Act Amendments of 1997 (IDEA 97) require that all students with disabilities, even those with the most significant disabilities, participate in state and district-wide assessment systems. Participation generally occurs in one of three ways:

with or without accommodations in the general, on-demand assessment, or through an alternate assessment. IDEA 97 requires that states report the performance of students with disabilities in the regular assessment and the alternate assessment with the same frequency and in the same detail that they report on the performance of non-disabled student [Section 612(a)(17)(B)(iii)]. Approaches to reporting that include both aggregation of all students with disabilities and non-disabled students together, as well as the disaggregated performance of students with disabilities are consistent with the requirements of IDEA 97 (Heumann & Warlick, 2000).

These requirements suggest that regardless of the nature of students' disabilities and nature of the alternate assessment used with the students, the alternate assessment results and those of the regular assessment program will need to be reported in a common fashion. Many states are opting to aggregate scores from the regular assessment and the alternate assessment (Thompson & Thurlow, 2001). The challenge in summing these assessment results is that the assessments, and the types of results each produces, are different. How can these assessments be reported together?

Types of Large-scale Assessment Programs

Most statewide assessments are one of two types: standards-based or norm-referenced (Olson, Jones, & Bond, 2001). Standards-based assessment programs directly measure the state's content standards, often using both multiple-choice and constructed-response items. Student performance is reported relative to the content standards of the state. The goal of these programs is to encourage all students to achieve all standards at high levels.

Norm-referenced tests are ones in which the performance of students is reported relative to a norm group (a representative group of students used as a comparison sample). Results are reported relative to this norm group, either in percentile ranks, normal-curve equivalents, grade levels, or other comparative scores. However, since the participation rate of students with disabilities in the norming samples is often low, the on-going participation rate of students with disabilities on such tests is also quite low (Thompson & Thurlow, 1999).

In this report, I focus primarily on how the results of standards-based assessments are reported. These results may be reported in several ways, including how students performed on each test item, how the student performed on a cluster of items measuring a content standard or a sub-unit thereof (e.g., benchmark or performance indicator), how the student performed overall on the assessment (raw score, scaled score, or other metric), and finally, the overall level of achievement, that is, which of several predetermined levels of performance the student's achievement fell into. States use three, four, or more levels to describe the performance of students. Terms such as "novice," "basic," "proficient," "meeting the standard," "advanced" or "exceeding the standard" may be used to describe the overall level of achievement of each student.

Types of Alternate Assessment

Because the small group of students who, due to the nature of their disabilities, require an alternate assessment, are quite diverse, the manner in which they are alternately assessed will also differ. Some common ways in which students are assessed include:

Checklists. This method relies on teachers to remember whether students are able to carry out certain activities. This technique has the advantage of permitting the rapid collection of information, but due to the nature of the observation, may not be highly reliable. Scores reported are usually the number of skills that the student was able to successfully perform. This method will permit the scores of students to be added up and reported.

Observation in Structured and Unstructured Settings. This assessment method encourages teachers, after training, to observe whether students are able to perform certain activities. Observation in unstructured situations is on-going observation of the student in everyday classroom and other settings, without any overt attempt to increase the likelihood that the skill will occur. By setting up structured situations, the teachers is setting up a structure in which the skill being observed is more likely to occur, thus making the observation of it more likely. Scores reported are usually the number of skills that the student was able to successfully perform. This method will permit the scores of students to be added up and reported.

Performance Assessments. These assessments are direct measures of the skill, usually in a one-on-one assessment. Due to the nature of students' disabilities, rarely are these paper-based assessments. More likely, the teacher and the student work through an assessment that uses manipulatives, and the teacher observes whether students are able to perform the assigned tasks. Such assessments have the disadvantage of being time-intensive, so that an assessment may be limited to only a handful of skills. Scores are typically assigned to each performance assessment, although in more complex performance assessments, there is an underlying scale of task complexity that may form the basis for reporting.

Samples of Student Work. Students may in the course of learning produce samples of work that demonstrate the skills being assessed. These "artifacts" can be assessed. While this assessment method has the advantage of using existing work in the assessment process, not all students will be able to produce samples, and even for those who do, it may not be possible to determine how much of the work is that of the student. Scores are assigned to each piece of work.

Portfolios. This assessment method uses a collection of student work, performance assessments, observations, and other data about students to judge student achievement. Usually, the various pieces collected to demonstrate the performance on each standard or group of standards are judged together, although occasionally, the entire content area or entire portfolio may be assigned a single score.

The Nature of Student Scores from Alternate Assessments

The nature of the scores derived from alternate assessments depends on the nature of the alternate assessment. Those that are comprised of checklists will yield multiple items scored 1-4 or 1-5, and an overall total score, which may be the number of items checked with 1s, 2s, 3s and 4s, or some other way of summarizing overall student performance. The same may be true for observational assessments, and even performance events, particularly when the latter consist of a number of small steps along a scale of performance, each scored on a 1-4 or 1-5 scale.

Portfolios, and some performance tasks, are scored using a scoring rubric, and the student score is derived from the nature and number of rubrics employed. In some assessments,

students' performances or portfolios are scored *holistically*, such as across multiple content areas using an overall performance rubric. In other cases, students' performances are scored *analytically*, that is, according to multiple dimensions.

The dimensions that states are using to score student work analytically focus on *student* performance, *program* opportunities provided to students, or a combination of the two (Thompson & Thurlow, 2001). The student dimensions can include a qualitative judgment about the overall level of performance of the student, as well as how close the student's performance was to the written content standards. The program dimensions could include whether students were provided instruction in multiple settings, whether they were provided opportunities to plan, monitor, and evaluate their work, whether they was evidence that they worked with non-disabled peers, and whether they were provided with appropriate human and technological supports.

These program opportunities can also be expressed as student performance dimensions as well: whether the student could demonstrate the skill in multiple settings; plan, monitor, and evaluate his or her work; work with non-disabled peers; and work independently (using appropriate human and technological supports). Depending on the nature of the scoring dimensions used, scores may be a simple sum across dimensions, be multiplied with one another, or be weighted in some fashion.

Reporting Alternate Assessment Information

While each state will report its alternate assessment information in somewhat different ways (in concert with differences in which their regular assessment program information is being reported), there are some similarities to these reporting schema. Depending on the nature of the regular assessment program – its purposes and methods – reports of the alternate assessment program may focus on *program* information or *individual student* information, or both.

Program data are reported in order to hold the school in which students are being taught, or the school that sent the student to the school in which the student is being taught, *accountable* for the student's performance. Programs might be held accountable for whether needed supports (human or technological) were made available, the inclusiveness of the student's program, the number of settings in which the student's accomplishments were evidenced, or the extent to which students are given opportunities to plan, monitor, and evaluate their own performances.

Individual student information is reported in order to describe the current level of the student's performance. It may focus on the qualities of the student's performance on the assessment, as well as how close to the content standards (as written) the student was able to come, the breadth or number of standards achieved, and the levels of supports needed to achieve at the level observed.

In almost all cases, the use of both student and program data is not for student accountability (to promote or graduate students, nor to retain them). Instead, it is to hold the school accountable for the learning opportunities afforded students, whether evidenced directly through student performance measures or more indirectly, through program measures.

While the parameters on reporting may vary from state to state, many states are opting to

aggregate the performance of all students with disabilities with the performance of students without disabilities, so that 100% of the students are counted (not simply “accounted for” by reporting them in a “no report” category). Since IDEA 97 requires that the performance of students with disabilities be disaggregated from the performance of other students, it has been suggested that the performance of the students with disabilities who take the tests with or without accommodations be added to the performance of students without disabilities and reported together (Heumann & Warlick, 2000).

The key question is how to accomplish these combined and disaggregated performances, when some students take a test and others participate in an alternate assessment comprised of a portfolio, performance events, or a checklist. Some sort of total report of results by content area is needed.

One way that states using portfolios accomplish this is by summing the performances of students to arrive at a total score. Once a series of total scores is determined, how the scores will be reported along with scores from the regular assessment program is determined by the manner in which these scores are labeled.

Score Scale for Reporting

States typically report the assessment results from their regular assessment program summarized at an overall test level, according to one of several performance descriptors. These performance descriptors or performance standards serve to describe “how good is good enough.” This helps to give additional meaning to the reports of results.

There are two ways to report the results. The first is to use an *absolute scale*, where the score scale used for the alternate assessment is equated to the score scale for the regular assessment program. This means that most of the alternate assessments will be reported in the bottom category of the regular score scale, that given the level of the performance and linkage to the standards as written, is viewed by some as an accurate representation of performance (Bechard, 2001). Nonetheless, to consign all of the students with significant disabilities to the bottom of the score scale also serves to reinforce low expectations for these students and perhaps to “punish” the educators who serve them. In the long run, this approach may discourage educators from offering a quality program or challenging the student to strive to accomplish more (since more challenging skills may lead to lower performance scores).

An alternative is to adopt the policy of reporting students in the alternate assessment on a *relative scale*. In this model, a score scale is constructed for students with significant disabilities in the alternate assessment, without equating the scale to the test scale. This can result in some students with significant disabilities being labeled “proficient” or “advanced,” even though their accomplishments are viewed by some as lower than students who took the test (Bechard, 2001). The result of using this scale is to reward educators who are offering a quality program in which students demonstrate significant accomplishments. This recognizes the successful work of educators, even when the nature of the student’s disability prevents the student from demonstrating typical performance levels. In the long run, this should encourage better programs for students. However, the “downside” of this is that the alternate assessment scale is different from the one used with the regular assessment program, and the difference may be interpreted as meaning “lower.”

Standard-setting Techniques

However the total score is derived (absolute or relative scales), the manner in which the performance of the student in the alternate assessment is categorized will depend on which of several standard-setting strategies is used. The strategies used for regular assessment programs may or may not be appropriate for standard-setting on the alternate assessment, depending on the regular assessment, the techniques used to set standards on it, and the nature of the alternate assessment component. Several techniques are used to set standards (Cizek, 2001) Each of these is described here and summarized in Table 1.

Table 1. Standard-setting Techniques that Might be Applied to Alternate Assessments

Technique	**Description**
Reasoned Judgment	A score scale (e.g., 32 points) is divided into a desired number of categories (e.g., 4) in some way (equally, larger in the middle, etc.); the categories are determined by a group of experts, policymakers, or others.
Contrasting Groups	Teachers separate students into groups based on their observations of the students in the classroom; the scores of the students are then calculated to determine where scores will be categorized in the future.
Modified Angoff	Raters estimate the percentage of students at the bottom score range who are expected to “pass” each test item; these individual estimates are summed to produce an overall percentage of items correct that correspond to the minimum passing score for that level.
Bookmarking or Item Mapping	Standard-setters mark the spot in a specially constructed test booklet (arranged in order of item difficulty) where a desired percentage of minimally proficient (or advanced) students would pass the item; or, standard-setters mark where the difference in performance of the proficient and advanced student on an exercise is a desired minimum percentage of students.
Body of Work	Reviewers examine all of the data for a student and use this information to place the student in one of the overall performance levels. Standard setters are given a set of papers that demonstrate the complete range of possible scores from low to high.
Judgmental Policy Capturing	Reviewers determines which of the various components of an overall assessment are more important than others, so that components or types of evidence are weighted.

Reasoned Judgment. The most straight-forward manner in which to set standards is for an appropriate group (either an expert panel, a representative group of users, or a policymaker group) to examine the score scale and to divide the full range of possible scores into the number of desired categories (Kingston, Kahl, Sweeney, & Bay, 2001). For example, a 32-point scale might be divided into 4 categories of approximately equal numbers of points (or different numbers of points in each of the categories), as the group sees fit.

The advantages of this strategy are that it takes little time, requires little in the way of a process, and does not hide the standard-setting in a cloak of mysterious statistical procedures. Presumably, the rationale for the choices is relatively evident. The major disadvantage is that rarely do natural divisions of performance occur, so that it may be difficult

to defend the choices that were made or the assignment of particular students to one level or another, since other reasonable people could arrive at different choices.

A special case of this technique that has been used with alternate assessments is to locate solid student exemplars for each score scale point. For example, if portfolios are scored on a four-point scale, the goal of this strategy is to locate solid 1s, 2s, 3s, and 4s of student work on all pertinent dimensions. These exemplars, which represent the different score levels of the scoring rubric, are then used for training purposes in scoring. A set of rules, which are predetermined, help determine the total score assigned to portfolios that are not given a consistent score across the various scoring dimensions.

Contrasting Groups. In this technique, a group of teachers familiar with the students, and with the definitions of the various groups into which students are to be placed, separate the students into these groups based on their observations of the students in their classroom (Livingston & Zeikey, 1982); then, the assessment scores in each of the groups are calculated. The distribution of scores among the different groups is examined; typically, where the scores between the two groups overlap is where the "cut score" between the two groups is set, since this is the point at which the classification errors are minimized.

A problem with this method is that it is highly dependent on the distributional characteristics of the sample. That is the rationale behind the development of a similar method called "Classroom Teacher Judgment" (Roeber, 2001).

The contrasting groups technique can be used with any type of assessment. One major disadvantage of it for alternate assessment is that teachers may not know what the performance of the student is on the types of skills (e.g., academic skills) measured by the alternate assessment. Nevertheless, it is a relatively easy technique to implement and is easily understood by educators and parents.

Modified Angoff. In this technique, an appropriate group (either an expert panel, a representative group of users, or a policymaker group) examines each test item in a multiple-choice exam. What each rater does is to estimate the percentage of students at the bottom of the score range (e.g., the "minimally proficient" or the "minimally advanced" students) who will be able to pass each test item. These individual estimates are then summed and result in an overall percentage of the items correct that correspond to the minimum passing score for that level of the test.

This technique is typically used with multiple-choice items, and is a classic standard-setting strategy for tests. It might be applied to checklists, such as those used to assess students with significant disabilities, although this use has not been tried. The major challenge to using the modified Angoff technique is a conceptual one: raters not only need to understand the theoretical "minimally-proficient student," they also have to determine how many of these students will pass the assessment. This is not an easy task, and hence one of the reasons why psychometricians looked for an improved way to set standards.

Bookmarking or Item Mapping. In this technique, an appropriate group (either an expert panel, a representative group of users, or a policymaker group) reviews a specially-constructed test booklet that is arranged in item difficulty order (Lewis, Mitzel, & Green, 1996). The standard-setter is asked to mark the spot in the booklet where a set percentage of minimally-proficient or minimally-advanced students would pass the item. An alternative method is for the standard-setter to mark where the difference in performance of the proficient

and advanced student on an exercise is a set minimum percentage of students.

This technique has the advantage of being usable with both multiple-choice and constructed-response exercises. It could be used with inventories or checklists since the percentage of students at each level in an inventory or checklist could be calculated. It would be challenging, however, to use this technique with portfolio assessments, where an overall score is derived.

Body of Work. In this technique, an appropriate group (an expert panel, a representative group of users, or a policymaker group) examines all of the data for a student and uses all of the information to place the student in one of the overall performance levels (Kingston et al., 2001). On a test, the multiple-choice and constructed-response performance of the student is examined together. Rather than examining test items, standard-setters examine students, and determine what combination of scores from the various test components would place a student in the advanced or proficient category. Standard-setters are given a set of papers that demonstrate the complete range of possible scores from low to high.

The advantage of this method is that all of the information about a student is used to set standards, which is an easier, more logical decision for a standard-setter to make. Discussions are more focused on tangible students rather than intangible percentages of students passing test items. This strategy could be used with checklists, inventories, and assessments using scoring rubrics (such as performance events or portfolios).

Judgmental Policy Capturing. In this technique, an appropriate group (either an expert panel, a representative group of users, or a policymaker group) reviews the various components of an overall assessment (which might be quite similar or quite dissimilar) and determines which of the components are more important than others. This might suggest weighting one type of item more important than another, or might be to weight one type of evidence (e.g., performance measures) as more important than another (e.g., a checklist) (Jaeger, 1994, 1995).

This method allows very dissimilar types of information to be used to make decisions about students, and permits these to be weighted differentially. This is not a technique that has been used widely, so little is known about its technical characteristics, particularly in student assessment.

Which Technique to Use with Which Alternate Assessments

The nature of the alternate assessment used will help determine the type of standard-setting procedure that will be used. In the case of portfolios, which include a variety of types of evidence, the arbitrary, preponderance of evidence, whole student, or policy-capturing procedures can be used. For alternate assessments that use performance events, with a range of indicators associated with them, any of the procedures could be used.

The technique used should take into account not only technical aspects of the alternate assessment strategies being used and the standard-setting strategy, but also the practical aspects of implementing the standard-setting technique for the alternate assessment process. For example, portfolios that take an hour to review may make the whole student procedure, while technically sound, impractical to implement on a statewide basis. Examining the amount of time that such reviews take – both in the beginning of such an effort and after reviewers are experienced – is an important part of the practical aspects that must be considered.

It will be important in the years to come to document the standard-setting approaches used with various types of alternate assessments. Unexpectedly high rates of students in the alternate assessment system who are achieving at high levels may be a reason for rethinking standard-setting approaches.

References

Angoff, W. M. (1971) Scales, norms, and equivalent scores. In R. L. Thorndike (ed.) *Educational measurement* (2nd edition, pp. 508-600). Washington, DC: American Council of Education.

Bechard, S. (2001). *Models for reporting the results of alternate assessments within state accountability systems* (Synthesis Report 39). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Cizek, G. J. (2001) *Setting performance standards: Concepts, methods, and perspectives*. Mahwah, NJ: Lawrence Erlbaum.

Heumann, J. E., & Warlick, K. R. (2000). *Questions and answers about provisions in the Individuals with Disabilities Education Act Amendments of 1997 related to students with disabilities and state and district-wide assessments* (Memorandum OSEP 00-24). Washington, DC: U.S. Department of Education, Office of Special Education and Rehabilitative Services.

Jaeger, R. M. (1994, April) *Setting standards through two-stage judgemental policy capturing.* Paper presented at the American Educational Research Association and the National Council on Measurement in Education.

Jaeger, R. M. (1995) Setting performance standards through two-stage judgmental policy capturing. *Applied Measurement in Education, 8*, 15-40.

Kingston, N., Kahl, S. R., Sweeney, K., & Bay, L. Setting performance standards using the body of work method. In G. J. Cizek (ed.), *Setting performance standards: Concepts, methods, and perspectives*. Mahwah, NJ: Lawrence Erlbaum.

Lewis, D. M., Mitzel, H. C., & Green, D. R. (1996, June) *Standard setting: A bookmark approach.* Paper presented at the Council of Chief State School Officers Large-Scale Assessment Conference, Colorado Springs, CO.

Livingston, S. A., & , Zeikey, M. J. (1982). *Passing scores: A manual for setting standards of performance on educational and occupational tests*. Princeton, NJ: Educational Testing Service.

Olson, J. F., Jones, I., & Bond, L. (2001). *State student assessment programs annual survey: Data on 1998-99 statewide student assessment programs*. Washington, DC: Council of Chief State School Officers.

Roeber, E. D. (2001). *Setting Standards Using Classroom Teacher Judgment.* Dover, New Hampshire: Measured Progress.

Thompson, S. J., & Thurlow, M. L. (2001). *2001 State special education outcomes: A report on state activities at the beginning of a new decade*. Minneapolis, MN: University of

Minnesota, National Center on Educational Outcomes.

Online Privacy Statement
This page was last updated on May 20, 2013

NCEO is supported primarily through a Cooperative Agreement (#H326G050007) with the Research to Practice Division, Office of Special Education Programs, U.S. Department of Education. Additional support for targeted projects, including those on LEP students, is provided by other federal and state agencies. Opinions expressed in this Web site do not necessarily reflect those of the U.S. Department of Education or Offices within it.